Программирование

Python

Работа с текстом

Кафедра ИВТ и ПМ
ЗабГУ

2018

План

Регулярные выражения

XML и HTML

docx

Обработка текстов на естественном языке

Ссылки и литература

Outline

Регулярные выражения

Регулярные выражения

Регулярные выражения (regular expressions) — формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов (wildcard characters).

Регулярные выражения

Для поиска используется строка-образец (**pattern**, «шаблоном», «маской»), состоящая из символов и метасимволов и задающая правило поиска.

Для манипуляций с текстом дополнительно задаётся строка замены, которая также может содержать в себе специальные символы.

Регулярные выражения

Краткая информация о способах задания шаблона приведена в справке по модулю.

```
help( re )
```

Регулярные выражения

```
import re

text = "The French invasion of Russia, known \
in Russia as the Patriotic War of 1812 and \
in France as the Russian Campaign, began \
on 24 June 1812 when Napoleon's Grande Armée \
crossed the Neman Rive"

res = re.search("\d\d\d\d", text)
if res:
    print("Найденная подстрока: " + res.group(0))
    print("Позиция подстроки: " + str(res.span()))
```

Найденная подстрока: 1812

Позиция подстроки: (71, 75)

Нечёткий поиск

```
import regex

# поиск подстроки amazing в строке amaging
# допустима одна или меньше опечатка ( e<=1 )
res = regex.match('(amazing){e<=1}', 'amaging')

if res:
    print("Найденная подстрока: " + res.group(0))
    print("Позиция подстроки: " + str(res.span()))
    print("Число опечаток, лишних вставок и недостающих символов: "
          + str(res.fuzzy_counts))
    print("Число позиций с опечатками, лишними вставками \
и недостающими символами: " + str(res.fuzzy_changes))
```

Найденная подстрока: amaging
Позиция подстроки: (0, 7)
Число опечаток, лишних вставок и недостающих символов: (1, 0, 0)
Число позиций с опечатками, лишними вставками и недостающими символами: ([3], [], [])

Нечёткий поиск

i<=3 permit at most 3 insertions, but no other types
d<=3 permit at most 3 deletions, but no other types
s<=3 permit at most 3 substitutions, but no other types
i<=1,s<=2 permit at most 1 insertion and at most 2 substitutions, but no deletions
e<=3 permit at most 3 errors
1<=e<=3 permit at least 1 and at most 3 errors
i<=2,d<=2,e<=3 permit at most 2 insertions, at most 2 deletions, at most 3 errors in total, but no substitutions

Outline

Outline

Outline

Outline

Ссылки и литература

- regex101.com - online regex tester

Ссылки и литература

Ссылка на слайды

github.com/VetrovSV/Programming